Reprinted from

PROCEEDINGS OF THE THIRD INTERNATIONAL CONGRESS OF

HUMAN GENETICS

THE UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS, U.S.A.
SEPTEMBER 5–10, 1966

PLENARY SESSIONS AND SYMPOSIA

Edited by
JAMES F. CROW
JAMES V. NEEL

THE JOHNS HOPKINS PRESS
BALTIMORE, MARYLAND 21218

CENSUS DATA FOR STUDIES OF GENETIC DEMOGRAPHY*

WALTER BODMER AND JOSHUA LEDERBERG

Department of Genetics, Stanford University School of Medicine, Palo Alto, California

INTRODUCTION

The formulation of population genetic models depends on the specification of three basic types of information with respect to the characteristics whose changes are to be described. These are firstly, the mode of inheritance of the characteristics, secondly, the mating patterns with respect to them, and thirdly, the differential selective forces associated with the characteristics. Very simple assumptions concerning these three types of information are usually made for the construction of population genetic models. This is mainly in order that they can be solved analytically. While the assumptions may often be reasonable and lead to models giving results which can easily be interpreted, information about man is available in much more detail than has generally been specified. Any attempts to construct comprehensive models of a human population will depend on the use of this detailed information.

Three major profiles of human populations are often left out of account by the simpler population genetic models: age, geography, and the complex profile of socioeconomic or behavioral characteristics which may have an important influence on reproductivity. The inheritance of socioeconomic characters defies any precise definition but can, at least in part, be described in terms of parent-offspring correlations. Mating patterns can be described in terms of socioeconomic and spatial

* We should like to thank William Buell for his assistance in the role of U.S. Census Bureau representative. We should also like to thank Dr. John Sved for very helpful suggestions and comments, and Dr. Howard Cann for his criticism of the manuscript. The work reported on is being supported by grants GM 10452 from the Institute of General Medical Sciences and Health and by grants from the Lt. Joseph P. Kennedy, Jr., Foundation and Syntex Laboratories, Inc.

parameters, and age. Selective differences are measured by the intrinsic rate of increase or "Malthusian parameter" [Fisher, 1930], originally defined by Lotka, whose calculation requires only a knowledge of age-specific birth and death rates. The *causes* of the differences, as opposed to their magnitudes, can, however, only be ascertained from a complete specification of all the factors affecting over-all reproductivity, such as age at marriage, probability of marriage, stability of marriage, fecundity, stillbirth, infant and fetal death rates, etc.

Demography in its broadest sense is simply the statistical study of human populations, but in its more usual, narrower, sense it refers to the study of mortality and fertility as a function of age, geographic and socioeconomic parameters. The proper measurement of selective differences depends on the tools of the demographer, as do also the descriptions of mating and migration patterns. Genetic demography can thus, perhaps, best be defined as the study of population genetic problems with the detailed approach of the demographer. Its pursuit depends on the availability on a large scale of demographic data collected with the *family* as a unit.

Large bodies of data are needed for two reasons. Firstly, to counteract sampling errors which confound the accurate measurement of small differences and, secondly, because of the small proportion of the total population which has any particular complex combination of socioeconomic and vital characteristics. Even the largest data files soon yield empty cells in complex cross-tabulations.

There are six basic types of sources of data:

1. population censuses,
2. vital statistics,
3. miscellaneous records (e.g., from hospitals),
4. special registries (e.g., registries of congenital malformations, such as exist in Scandinavian countries and such as the registry of handicapped children in British Columbia, Canada),
5. parish books,
6. special purpose surveys.

Each of these sources has its advantages and disadvantages. Thus, while much more detailed and accurate information can be obtained from special purpose surveys than from censuses or vital statistics, it may be hard to avoid sampling biases, and the cost of an adequately large survey may be prohibitive. Vital statistics provide relatively accurate but very limited information on almost the entire population, while censuses provide more information on a similar scale, but are probably more subject to response errors and sampling biases and refer only to a single time-point in the history of a population. Records from these sources do not relate to the family as a unit without cross reference to other sources of information. All of them can, however, be of use, especially if they can be related through the process of "record linking," strongly advocated by Newcombe and his colleagues.

The study of genetic demography requires the use of computers at three different stages. In the first place they are essential for the basic handling, organization, and tabulation of data. The collation of different records for record linking is a costly and time consuming process which will be greatly facilitated by the availability of large, fast-access memories and time-sharing facilities, allowing multiple simultaneous access to the same large body of data. At the second stage, computers are essential for

the statistical analysis of large series of cross-tabulations. Finally, the interpretation of the results depends on studying the properties of complex population genetic models which mostly can only be revealed by comprehensive simulation, using input parameters specified by the data.

In this paper we shall discuss the uses and limitations of census data for studies of genetic demography. The discussion is based on our experience in analyzing the 5 per cent sample of the United States 1960 Population Census. This work is being carried out in close collaboration with the U.S. Bureau of the Census. Technical operations in programming the data-processing are carried out at Stanford under the supervision of a representative seconded from the Bureau. This has proved to be a very satisfactory arrangement as it reconciles the need for:

a) control of confidentiality of the files,
b) the Bureau's interest in the analysis,
c) our access to the summary data,
d) supervision of data reduction and the preparation of summary files,
e) efficient communication with the Bureau on technical details.

The initial output of the study will be a report on child-spacing planned by the Bureau as a volume in its regular series on the 1960 census, while the long-term aim is in the comprehensive analysis of reproductivity, migration, and mating patterns in relation to age, socioeconomic, and other available parameters.

Information Available from Census Data

The starting point for our analysis is a reduced version of the original 5 per cent file, from which information not relevant to our particular interests has been excluded. The 5 per cent file is a machine-selected subsample of one-fifth of the original 25 per cent sample schedules. The 25 per cent sample is derived from a random sample of every fourth household from which more detailed information was collected than that required for the complete (100 per cent) census. The basic sampling unit of the census is the household. In group quarters (such as homes for the aged, deaf, and blind, schools for delinquents, etc.) the 25 per cent sample consisted of every fourth person, in the order they were listed within the group quarters.

For convenience of data-handling, the file has been classified into the following sub-files:

1. Single family, husband-wife, households with one or more children (all of whom were present at the time of the census, and so included with their parents in the schedule collected from the household).

2–4. Other single-family households in which not all the children ever born to the wife or female head of the household were present at the time of the census and in which she has:

a) none or one child ever born,
b) two or three children ever born,
c) four or more children ever born,

5. Multiple family households.
6. Persons in nonfamily households or group quarters.
7. Miscellaneous anomalous households (e.g., no head of household, or data

inconsistency, such as number of children present exceeds reported number of children ever born).

8. Multiple births. This file contains a duplicate of all households from files 1, 3, 4, and 5 which contain instances of multiple births as indicated by the birth dates of the children present in the household at the time of the census.

The data in each file are sequenced by state within geographic regions (north, south, east, or west). Individuals and secondary families within each household are arranged in a definite sequence according to their relationship to the head of the household, while the children within a family (primary or secondary) are sequenced by their age, with the oldest first. Only the first file includes simple complete families. It is the most convenient for analysis, though, as will be discussed below, it certainly represents a biased selection of the population. Files 2 to 5 include incomplete families, that is, households in which not all the children ever born to the wife or female head were still present at the time of the census. These pose special problems which, also, will be discussed in some detail later. File 7 simply isolates the small number of anomalous households, and can probably be largely ignored, while file 8 is a special purpose file for the analysis of data on families containing multiple births.

A summary of the data items included for each household and each person within a household is given in Table 1. The items are divided into three groups according to whether they concern the whole household, individuals of all ages, or only individuals over fourteen, that is, adults. The types of data can be divided into five categories as follows:

a) *Socioeconomic parameters.* These include all the household characteristics which define the type of residential area, quality of the dwelling, income and occupation of the head of the household, and the socioeconomic scores and consistency. Additional information from the other two groups includes schooling, veteran status, details of employment, as well as individual income and occupation. Many of these parameters are, of course, highly correlated.

b) *Geographic parameters.* These include the birthplace of an individual and of his parents, as well as the mother tongue (if foreign born), and mobility.

c) *Special individual characteristics,* namely sex, race, and birth date.

d) *Reproductivity* as defined by the number of times married, marital status, number of children ever born and, in many cases, the parental age at childbirth. This latter information is particularly important for the determination of age-specific birth rates.

e) *Relationship to head of household and family relationship.* This information is the basis for the construction of family groupings within households. It is essential for the determination of parental age at childbirth and for the identification of husband-wife pairs for the study of marriage patterns.

The census provides information on only two of the basic types of data needed for the construction of comprehensive population genetic models, namely, marriage and fertility patterns. Since, in this country, mortality rates are now very low, the fertility data should account for the major fraction of any differences in over-all reproductivity. However, the collection of data by the household rather than the family, means that much of the fertility data is incomplete, since the birth dates of children no longer in the same household as their parents are unavailable. The problem this poses for the determination of over-all fertility patterns will be discussed later in more detail.

TABLE 1. DATA ITEMS INCLUDED IN THE SPECIAL CLASSIFIED 5 PER CENT CENSUS FILE

Household characters	Characteristics of all persons	Characteristics of persons 14 years old, and over
1. Type of residential area (rural farm or nonfarm, urban) and size of urbanized area	1. Sex	1. Times married
2. Size of city or place	2. Race	2. Marital status
3. Size of standard metropolitan statistical area	3. Relation to head of household	3. Total children ever born
4. Tenure (own, rent, or group quarters)	4. Family relationship	4. Veteran status
5. Value of owner-occupied unit	5. Date of birth	5. Date of first marriage
6. Rent in rented unit	6. School years completed	6. Date last worked (if not worked in week prior to the census)
7. Socioeconomic status	7. Current school enrollment status	7. Hours worked in week prior to the census
8. Socioeconomic consistency	8. Nativity and parentage (native or foreign born, parents native or foreign born)	8. Employment status
9. Income of primary family	9. Mobility (born in same or different state)	9. Weeks worked in 1959
10. Industry of primary family head or individual	10. Place of birth for native born (state)	10. Total person's income
11. Type of group quarters	11. Parents' place of birth	11. Occupation
	12. Place of birth for foreign born (country)	
	13. Mother tongue of foreign born	

Notes:

(a) Standard Metropolitan Statistical Areas are regions surrounding major centers of population concentration, which have been defined by the Census Bureau as a guide to the definition of urban areas.

(b) The socioeconomic status score is the mean of three numerical scores assigned to income, occupation, and education. Sociocconomic consistency is a measure of the maximum difference between these three scores.

(c) Relation to head of household and family relationship between them identify all possible relationships between individuals within a household, in particular the family units.

(d) All dates are given in calendar quarters.

For further details concerning some of these parameters and also the general procedures followed for the 1960 Census consult U.S. Bureau of Census [1964a, b]. A 1/1000 sample of the 1960 U.S. Population Census is available to qualified investigators. The accompanying description and technical documentation contains a detailed statement of the information available from this census (U.S. Censuses of population and housing: 1960, 1/1000, 1/10,000. Two national samples of the United States. Description and technical documentation).

Perhaps the most severe basic deficiency of the data from a genetic point of view is the complete lack of information relating two generations, and hence the lack of any information for determining the relative contribution of genetic factors to the available socioeconomic parameters. Only in so far as there are any inherited (social or genetic inheritance) components of socioeconomic characteristics, will the mating and reproductivity patterns influence the future composition of the population with respect to them. Nevertheless, reproductivity differences with respect to these socioeconomic parameters will set an upper limit to the possibilities for selection, as has been emphasized by Crow [1958].

GENERAL APPROACHES TO THE ANALYSIS OF THE DATA

The basic problem is the analysis of fertility and mating patterns as a function of the available data categories. A data category is defined by a particular cell in some, possibly complex, cross-tabulation. We may, for example, be interested in a

classification of mating pairs by school years completed, nativity and age at first marriage. Alternatively, we may wish to determine age-specific birth rates as a function of the school years completed by husband and wife, and whether or not one or both of them are foreign born. The maximum practical level of complexity for the definition of data categories is still hard to predict at this preliminary stage of our analysis. Severe limitations will be imposed both by the amount of data which is available and the level of comprehension which can be achieved for complex cross-tabulations. The number of combinations which are potentially of interest is almost unlimited.

Mating patterns are defined by a two-way correlation table giving the relative frequencies of all possible mating pairs with respect to the relevant data categories. Migration patterns, similarly tabulated, may be defined by the probability that an individual of given characteristic migrates from one given state to another. These probabilities can be used as input parameters for models describing changes in population constitution both in space and time.

Fertility patterns are not so easily defined. The intrinsic rate of population increase, r, which is the basis for the proper measurement of selective differences, is the solution of the equation

$$\sum_x e^{-rx} l_x b_x = 1 \tag{1}$$

where x measures time in units of one or more years. b_x, the age specific birth rate, is one-half the number of births to individuals of ages x to $x + l$, l_x is the probability of survival from birth to age x, and summation is over all ages. Within a specified data category, the age-specific birth rate for any given age group is one half the number of births produced at age x divided by the number of relevant individuals and is readily obtained from complete family data. Assuming the age-specific birth rate of survivors to the given age of a group is not different from that of those who died before the census, the observed rate will be representative of the whole data category. In the United States the reduction of maternal mortality has been one of the most spectacular achievements of medical progress, and the proportion of individuals who die during the reproductive years is, in any case, very small. This means, as already pointed out, that differential mortality at this stage of life is likely to be of minor importance as a factor determining differences in the intrinsic rate of increase. Differences in earlier mortality, especially infant mortality, may be more important, but no relevant information is available from the census as presently conducted. The intrinsic rate of increase for a given data category can be calculated from equation (1) using values of l_x obtained from standard life tables constructed from other sources than the census, usually vital statistics. The number of children born per person born is the net reproductive rate R_o, given by

$$R_o = \sum_x l_x b_x \tag{2}$$

The net reproductive rate (R_o), intrinsic rate of increase (r) and generation length T are related by the equation

$$r = \frac{\log_e R_o}{T} \tag{3}$$

Thus, in the absence of differences in generation length, relative selective differences

can be adequately measured by the net reproduction rate. If mortality during the reproductive years is negligible, then

$$R_o = l\,B \tag{4}$$

where l is the probability of surviving to the reproductive age and B the total number of births per person for people who have just reached the end of their reproductive years. The birth rate as a function of age is now required only for the detection of that component of selective differences due to variations in generation length.

The determination of the causes of a difference in reproductivity, as opposed to the measurement of its magnitude, requires a more complete description of the factors which affect reproductivity. The only ones available from census data are the probability and stability of marriage as a function of age, and the birth patterns as a function of age, the latter defined by the distributions of the intervals between births and the age at marriage. Many factors, such as the monthly probability of conception, birth control practice, fetal death and stillbirth rates, and the length of the post-partum sterile period influence birth intervals. In the absence of any information concerning these factors separately, birth-interval data cannot give much indication as to the cause of a fertility difference. Thus the only causes of a reproductivity difference which can be reasonably identified by census data relate to marriage patterns and their stability as a function of age, and the distribution of the age at childbirth.

It has already been emphasized that one of the goals of a study such as this one is to provide some of the input parameters needed for the construction of models to describe temporal and spatial changes in population structure with respect to socio-economic and other parameters. Much of the relevant input data will be in the form of distributions, such as those needed to describe the age at marriage, the difference in age between husband and wife, family size, interval between births, inter-generational migration distances, and so on. The computer simulation of models is greatly facilitated by the use of appropriate theoretical distributions which can be fitted to the various observed distributions. Thus, for example, family size can usually be described by a negative binomial or a modified negative binomial distribution [Brass, 1958], and time intervals can often be fitted by gamma distributions. In many cases, the effects of differences in a given distribution can also be conveniently studied in terms of the parameters of the appropriate theoretical distribution.

Some special problems, not related directly to the determination of mating and fertility patterns, can also be studied with census data. Of particular interest to us are the investigation of seasonal differences in birth rates as a function of socio-economic parameters, the study of the sequences of sexes within families, and the characteristics of families including multiple births.

Problems and Deficiencies of Census Data

The general deficiencies of census data for studies in genetic demography have already been emphasized. They are mainly, the lack of mortality and morbidity data, the lack of data relating two generations, and the fact that the data are collected by the household rather than the complete family. Any large social survey, in par-

ticular a census, is subject to errors of response and deficiencies in the completeness of enumeration. In this section we shall review the incidence of missing data items and the techniques used to compensate for them and also for the incompleteness of much of the family data.

The 1960 U.S. Census schedule was a special form designed for microfilming in such a way that information from the microfilm could be transferred directly to magnetic tape (by FOSDIC—Film Optical Sensing Device). Nevertheless, some clerical editing of the schedules was necessary, in particular, numerical coding of written entries. Intensive quality-control checks were made at this stage, which is, therefore, unlikely to be a significant source of error. Some household schedules proved to be "unreadable." These were cancelled and replaced by household schedules from the same area which had similar characteristics with respect to certain major parameters. These latter were, thus, replicated once to replace a cancelled household. A similar adjustment was made for a slight bias toward larger households in the 25 per cent sample. The total number of households in the 25 per cent sample was 13,005,524 and the corresponding total number of persons 49,319,625. The proportions of households replicated because of schedule "unreadability" and size bias were 0.98 and 0.22 per cent, respectively. These, also, are therefore unlikely to be important sources of error.

The major source of error is due to nonresponses on the original schedules. The complete count (100 per cent survey) contained questions on relationship, age, sex, race, and marital status. Information on an individual was only accepted for further processing if at least two of these items contained entries, one of them being relationship, sex, or race. Missing or inconsistent entries in an accepted person's data were "allocated" by computer according to the following general procedure. As a magnetic tape was processed a running record was maintained of the five complete count characteristics for a certain number of individuals nearest to the one currently being processed. This record was updated, one person at a time, as each individual with responses in all five items was encountered. A nonresponse was allocated by replicating the item from the nearest previous reported person in the running record who had the same five characteristics (or four, if the item was one of these five). Since records are arranged according to contiguous regions within a state, the allocated items generally come from individuals in nearby regions. This procedure is equivalent to an empirical stratified sample based on geographic area and the five complete count characteristics. Almost all tabulations published by the census are based on data which includes allocations.

In most cases, each allocation in an item was "flagged" in a special allocation bit position of the computer record. Allocated items can therefore be identified, their rates determined as a function of available parameters, and the characteristics of their distributions compared with nonallocated items. This opportunity provides a very important control over the rates of errors and biases which might be caused by nonresponses.

Nonresponse rates vary considerably according to the type of question and the socioeconomic attributes of the respondent [U.S. Bureau of Census, 1964a, pp. 342–44, and 1964b, pp. 317–23]. The over-all per cent of persons who were acceptable (namely, had entries in at least two of the complete count characteristics) was 98.5, while the per cent of individuals with one or more allocated nonresponses was 18.9. Over-all allocation rates for relationship, sex, birthdate, and marital status

were 0.7, 0.2, 1.0, and 0.6 per cent, respectively. The effect of allocation for these items can probably be ignored with impunity. Allocation rates for other items of interest ranged from 8.3 per cent for school enrollment, to 4.9 per cent for highest school grade completed. The allocation rates are generally highest in central city areas and lowest in rural areas, with maximum differences (between the general types of area) of about 50 per cent. The over-all rate for children ever born is 6 per cent, and ranges from 12.5 per cent for the District of Columbia to 3.5 per cent for Montana.

As might be expected, the allocation rate for children ever born varies somewhat according to age, being 14.1 per cent for ages 15–19 years, 6.8 per cent for 20–24, but ranging only between 4.8 and 5.5 per cent in the five 5-year intervals from ages 25 to 49. All the rates are appreciably higher for nonwhites than for whites. There is no doubt that the allocation rate for all items varies inversely with socioeconomic status and must be taken into account in any valid comparisons of reproductive performance between defined data categories. Continual monitoring of the consistency and quality of the data is an essential and costly feature of the analysis of such large bodies of data as are obtained from the census.

Undoubtedly, the major problem in the use of census data for determining age-specific birth rates is the fact that much of the data on birthdates of children within a family is incomplete because of the use of the household as the sampling unit. Information on children who have left the household is missed, though the total number of children ever born to each female is, of course, recorded. There are many socioeconomic, cultural, and other factors (including, for example, mental retardation) which are correlated with the age at which children leave the parental household. The average characteristics of families with all children present may, therefore, differ markedly from those with one or more children absent. It will, in practice, be impossible to control (or even determine) all the variables which may influence such a bias. There may also be inherent ascertainment biases in the birth-interval distributions obtained from such household data, which are not due to socioeconomic and other stratifications. Thus if, for example, the age of a child were the main factor determining when it leaves a household, it can be shown that families from older married couples with all children ever born present would, in general, be biased toward longer birth intervals.

The longer a couple has been married, the older their children and so the higher the probability that one or more of them will have left the home. There is, thus, a basic conflict between obtaining data on complete families as opposed to data on couples whose fertility is completed. Only complete families whose fertility is completed provide really satisfactory data for determining age-specific birth rates.

There is an obvious need to obtain some information on the characteristics of the birthdate and interval distributions for families with children missing from the household. The census bureau has suggested procedures for "allocating" information on missing children on tables constructed from sample surveys (August, 1959, Current Population Survey) which specifically provide information on these distributions for children absent from home. Such procedures suffer from three major drawbacks:

1. The allocation distributions are based mostly on subclassification only by race and marital status. A bias may thus be introduced when these allocation distributions are used for more complex cross-classifications, such as will be required for our analysis of the 5 per cent sample of the 1960 U.S. Population Census.

2. Many of the allocations are based on small numbers leading, possibly, to relatively large sampling errors when they are applied to a considerably larger body of data.

3. There may be a bias in the source data for the allocation tables, for example, if these changed from the 1959 CPS to the 1960 census. The allocation procedure effectively combines data from complete families with the allocation distributions. The confounding of these two sources of information may considerably dilute the value of the actual data from complete families.

A general procedure has been devised for the estimation of birth intervals from families with children absent from home, using only the information provided in the 1960 population census. Knowledge of birth-interval distributions would also facilitate the calculation of age-specific birth rates. The estimation problem is considered for a given data category, for which child-spacing distributions are desired. Age at first marriage and age at the time of the census will certainly be amongst the most important parameters defining any data category. To illustrate the method, we consider the case of families with two children ever born. As indicated in Table 2, there are four types of families according to which child is present or absent. Families of types 2 and 3 with one child missing (either the first or the second), cannot be distinguished from the census data as collected. We thus observe directly only the quantities p_1, $p_2 + p_3$, p_4, the distribution of the interval from marriage to the birth of the first child and from the first child to the birth of the second child for completed (type 1), and the distribution of the interval from marriage to the birth of the child which is present for families of types 2 and 3. Our concern in trying to obtain some estimate of the interval distributions for families of types 2 and 3, with one child missing, is in case the distributions differ significantly from those observed for the completed families. If this were the case, then these differences would have to be taken into account in describing the interval distributions for the particular category under consideration. It is, of course, clear that no information can be obtained on the interval distribution for families of type 4, where both children are missing. Possible biases introduced by using only data from the completed families will, of course, be minimized when the proportions p_2, p_3, and p_4 are small.

The observed distribution of the interval from marriage to the child present for families with one child missing will be a weighted combination of the distribution of the interval from marriage to the birth of the first child for families of type 2, and the distribution of the interval from marriage to the birth of the second child for families of type 3. The latter is, in fact, the distribution of a "double" interval which, of course, would be expected to be appreciably longer than the corresponding distri-

TABLE 2. TYPES OF FAMILIES WITH TWO CHILDREN EVER BORN ACCORDING TO WHICH CHILDREN WERE PRESENT AT THE TIME OF THE CENSUS

Family type	1st child	2nd child	Proportion of families of given type
1	+	+	p_1
2	+	0	p_2
3	0	+	p_3
4	0	0	p_4

+ = child present
0 = child absent

$p_1 + p_2 + p_3 + p_4 = 1$

bution for a single interval. The observed distribution should therefore be bi-modal and its components resolvable by fitting a weighted mixture of two distributions describing in turn the expected distributions for the interval from marriage to the birth of the first child for families of type 2 and the interval from marriage to the birth of the second child for families of type 3. More specifically if $f_2(x)$ represents the expected distribution of the interval from marriage to the first child for families of type 2 and $f_3(x)$ represents the probability density function for the distribution of the interval from marriage to a second child for families of type 3, then the expected probability density function for the interval from marriage to the birth of the child present for families with one child absent is given by

$$\frac{p_2}{p_2 + p_3} f_2(x) + \frac{p_3}{p_2 + p_3} f_3(x).$$

Given a theoretical distributional form for birth intervals, i.e., for $f_2(x)$ and $f_3(x)$, we can use standard statistical procedures, such as maximum likelihood, to fit this expected mixed distribution to the observed distribution. This will give estimates of the proportions p_2 and p_3 and of the parameters defining the distributions $f_2(x)$ and $f_3(x)$. We can then ask the question as to whether these distributions differ significantly from the corresponding distributions observed for completed families (type 1) and so assess biases introduced by ignoring incomplete families. The general approach outlined above can easily be extended to the larger families, though the hope of extracting useful information from larger families with more than one child missing from the household is clearly limited.

It is anticipated that some general two or three parameter distribution will be found which has a suitable analytical form for fitting birth-interval distributions. The gamma distribution gives a very poor fit to data on the interval from marriage to first birth, probably because of its inability to take account of the large mode at nine to ten months, together with the long tail of the observed distribution. It gives a somewhat better, but still inadequate, fit to data on subsequent birth intervals. Models constructed by Perrin and Sheps [1964] and others, show that the single birth interval can itself best be represented by a mixture of at least two distributions, though we have not yet been able to find any convenient representation for a suitable mixture. A serious limitation to this general approach is the difficulty of resolving mixed distributions, given a limited number of observations. Experience only will show how satisfactory the method is for any given level of cross-classification.

Discussion—Prospects for the Future

Population projections are an essential part of both population genetics and demography, as well as providing much basic data for the planning of our social and economic future. Their pursuit requires a knowledge of demographic parameters as a function of genetic relationships or, in other words, the collection of demographic data on a large scale, with the family as the unit. Population censuses go some way toward fulfilling this need though, as we have discussed, they suffer from some very severe deficiencies. Their very availability on such a large scale, however, challenges us with the problem of their analysis in the hope, at least, that we can learn from past experience what must be done in the future.

The long-term answer to these major problems of data collection undoubtedly lies in the comprehensive computer-aided linking of records from different sources, through some basic identifier assigned at or referable to the birth event. Useful information can still, however, be collected from censuses. Simple changes, such as the specification of dates in months instead of quarters, could be of great value. Smaller sample censuses, perhaps at a one or two per cent level, could be designed with more comprehensive questionnaires and still provide data on a large enough scale for most of the needs of a genetic demographer. It would be relatively easy to include questions on the birthdates of children missing from the household at the time of the census, and so complete the family. In addition, simple questions concerning the brothers and sisters of heads of households and their spouses, providing at least some information relating two generations, could also be included. This is a time when lobbying by biologists, including geneticists and other health-oriented research workers, for the collection of more data appropriate to their interests, could be of immense value to them and the society they live in. Our citizens and their representatives are properly apprehensive about the potential intrusions on their privacy represented by linkable vital records. Public acceptance of measures like the assignment of identifying numbers will require a substantial campaign of education and legislation. This must show the great importance of population studies for human welfare. It must also sustain rugged legal sanctions to assure personal privacy against the temptations opened to individual abuse by the very existence of the same data needed for statistical knowledge.

SUMMARY

The general aims and data requirements for studies in genetic demography are reviewed. The main opportunities in the analysis of census data lie in the study of mating, migration, and fertility patterns as a function of age, socioeconomic, and other available parameters.

A brief outline is given of the general content of census data, as obtained from the 5 per cent sample of the 1960 U.S. Population Census, insofar as it relates to genetic demography. The over-all approach to the analysis of the data, emphasizing especially the determination of age-specific birth rates, is also reviewed. The major limitations of the data are that they relate only to a single generation and do not, in general, provide data on complete families. Nonresponse rates for individual data items vary considerably with socioeconomic status and must be carefully controlled for any valid comparison of fertility differences between data categories defined by more or less complex cross-tabulations. An approach to the estimation of birthdates for "missing" children is outlined. There is an urgent need for the collection of more and more appropriate data.

REFERENCES

BRASS, W. 1958. Models of birth distributions in human populations. *Bull. de L'Inst. Int. de Stat.* **36:** 165–78.

CROW, J. 1958. Some possibilities for measuring selection intensities in man. *Hum. Biol.* **30:** 1–13.

FISHER, R. A. 1930. The genetical theory of natural selection. London: Oxford Univ. Press (2nd ed., 1958, Dover Publications).

PERRIN, E. B. & SHEPS, M. C. 1964. Human reproduction: A stochastic process. *Biometrics* **20:** 28–45.

U.S. Bureau of the Census. 1964a. *U.S. census of population: 1960.* Vol. 1. *Characteristics of the population,* Part 1. United States Summary. Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census. 1964b. *U.S. census of population: 1960. Subject reports. Women by number of children ever born.* Final report PC(2)-3A. Washington, D.C.: U.S. Government Printing Office.